Deep Ensemble Learning for Skin Lesion Classification from Dermoscopic Images

Ahmed H. Shahin
Medical Imaging and Image Processing Group
Center for Informatics Sceinces
Nile University
Giza, Egypt
a.hassaan@nu.edu.eg

Ahmed Kamal
Medical Imaging and Image Processing Group
Center for Informatics Sciences
Nile University
Giza, Egypt
ahm.kamal@nu.edu.eg

Mustafa A. Elattar
Medical Imaging and Image Processing Group
Center for Informatics Sciences
Nile University
Giza, Egypt
melattar@nu.edu.eg

***Abstract*— Skin cancer is one of the leading causes of death globally. Early diagnosis of skin lesion significantly increases the prevalence of recovery. Automatic classification of the skin lesion is a challenging task to provide clinicians with the ability to differentiate between different kind of lesion categories and recommend the suitable treatment. Recently, Deep Convolutional Neural Networks have achieved tremendous success in many machine learning applications and have shown an outstanding performance in various computer-assisted diagnosis applications. Our goal is to develop an automated framework that efficiently performs a reliable automatic lesion classification to seven skin lesion types. In this work, we propose a deep neural network-based framework that follows an ensemble approach by combining ResNet-50 and Inception V3 architectures to classify the seven different skin lesion types. Experimental validation results have achieved accurate classification with an assuring validation accuracy up to 0.899.**

Keywords—skin lesions, diagnosis, melanoma, deep neural networks, deep learning.

I. INTRODUCTION

Melanoma accounts for 1% of all skin cancer cases, but for the vast majority of its deaths. More than 90,000 new cases of melanoma are expected to be diagnosed in the US in 2018 [1]. Early detection and diagnosis of skin cancer are the most effective ways towards faster and more successful treatment. Dermatologists often adopt different diagnostic guides such as ABCD rule (asymmetry [A], irregularity of borders [B], unevenness of distribution of color [C], and diameter [D]) to identify melanoma lesions [2]. In practice, dermatologists can diagnose the lesion to only two classes (Benign and malignant). However, it is quite challenging to diagnose extra subcategories based on visual features as shown in Fig. 1.

Computer-assisted tools have shown an outstanding performance in the clinical practice and can be used as an efficient physician assistant for diagnosis. Recently, deep convolutional neural networks (DCNNs) [3] were used to detect and/or classify lesions based on dermoscopic and non-dermoscopic images of skin [4]. In terms of accuracy, DCNNs outperformed highly qualified dermatologists [5]. DCNNs classify the lesions based on high-level features rather than the conventional method incorporating the low-level dermoscopic visual information that requires a segmentation step beforehand to extract those features. Two well-known deep learning classification approaches (ResNet [6], and Inception V3 [7]) have shown an outstanding performance in other image classification challenges such as visual object recognition [8].

In this work, we propose a challenging solution for lesion classification problem by providing an automated approach that efficiently performs a reliable disease classification with the ability to differentiate between seven different types of lesions: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis / Bowen's disease (intraepithelial carcinoma) (AKIEC), benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL), dermatofibroma (DF), and vascular lesion (VASC). We investigate the idea of lesion diagnosis based on ensembles of two widely-used CNN architectures – ResNet and Inception V3.

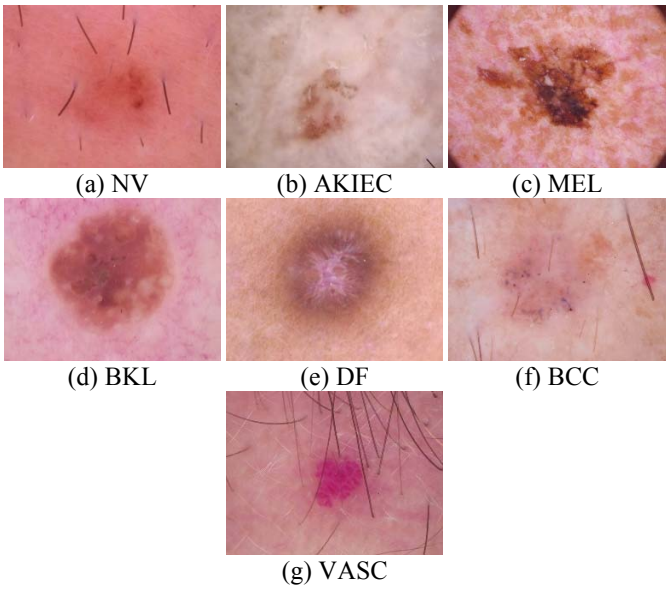

Fig. 1. Lesion images with different lesions states

II. RELATED WORK

Traditionally, a melanoma classification framework was proposed by Barata et al. that was based on feature extraction and classification modules [9]. Handcrafted features such as color histogram, edge histogram, and a multi-scale variant of color local binary patterns are extracted from the input image [10]. After feature extraction step, machine learning methods such as support vector machines, k-nearest-neighbors, logistic regression, and neural networks are applied to perform the classification task. All of those methods have yielded suboptimal accuracies and suffered from instability of results [11].

Codella et al. proposed an algorithm that combined the deep learning, sparse coding, and SVM for melanoma classification [12]. Their approach showed an acceptable performance on dermoscopic images while classification into only three classes (melanoma, atypical nevi, and benign lesions). In [10], an improvement over their previous

978-1-5386-8154-1/18/$31.00 ©2018 IEEE

method has been introduced by incorporating multi-contextual analysis of the extracted features to perform the classification task with moderate success.

DCNNs behave as an automatic feature extractors from dermal images in an iterative way, that yields to the promotion of the most discriminative features in the image. Afterwards, convolutional layers are followed by fully connected layers that perform the classification task based on the extracted features. Lopez et al. proposed a deep learning approach for skin lesion classification to melanoma/benign lesion [13]. They used the VGG network, which is a CNN architecture that had an outstanding performance in other visual classification tasks. They investigated the idea of lesion classification to melanoma/benign on ISIC 2016 dataset [14] using transfer learning. However, VGG has performed poorly in comparison with more complicated and modern architectures such as ResNet, and Google Inception in classification tasks.

III. Methods

In this paper, we introduce a new approach for skin lesion classification using ensemble approach by combining two deep learning architectures (ResNet and Inception V3). Our approach has been implemented, validated, and benchmarked against the latest publicly available dataset for dermoscopic images [9]. We use ResNet and Inception V3 to classify the suspected lesions to seven different classes of the disease.

A. Dataset

Our dataset was obtained from the "ISIC 2018: Skin Lesions Analysis Towards Melanoma Detection" challenge [15], [16]. The dataset consisted of 10,015 RGB dermoscopic images and their corresponding ground truth labels. The input data are dermoscopic lesion images in JPEG format. The lesion images were acquired from different anatomic sites, and from several different institutions. The images have a unique resolution of 450x600. The ground truth indicates the diagnosis of each input lesion image to seven possible types of skin lesions (MEL, NV, BCC, AKIEC, BKL, DF, and VASC). The provided ground truth images have been confirmed by one of these methods: histopathology, reflectance confocal microscopy, lesion follow up over two years, or consensus of three or more expert dermatologists.

B. Neural Network Archetictures

In this study, we investigate the performance of ResNet and Inception architectures separately and their ensemble (the combined architectures) for skin lesion classification. Generally speaking, ResNet and Inception both have the same concept of extracting the features in the convolutional layers then the high-level reasoning in the neural network and the classification step is done using the fully connected layers. We have modified the fully connected layers in both architectures to output seven classes instead of 1000 classes as was proposed for ImageNet.

1) ResNet Archeticture

ResNet has been introduced in 2015 by Microsoft for visual recognition tasks [6]. It introduces the idea of residual connections (see Fig. 2) as a solution for the problems of accuracy saturation and degradation when increasing the network depth. ResNet comes with different variants such as: ResNet-50, ResNet-101, and ResNet-152 according to number of layers used. In this work, ResNet-50 was selected to be used for our lesion classification problem due to the simplicity of the skin lesion image where only a single object is in the scene. The input image is passed through 7x7 convolutional layer, followed by pooling layer, stack of residual blocks, average pooling layer, and finally three fully connected layers as a classifier that outputs the probabilities of image correspondence to the seven classes. Each residual block consists of two 3x3 convolutional layers and every convolutional layer is followed by ReLU as an activation function.

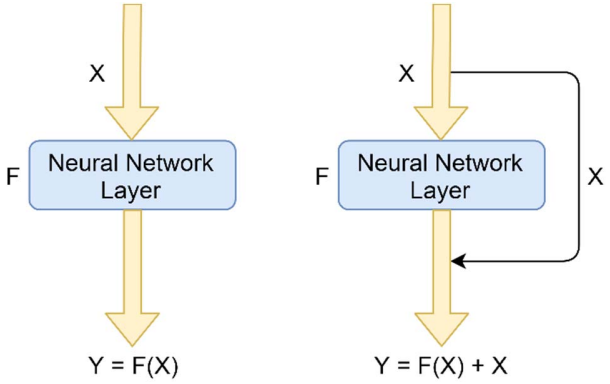

Fig. 2. An illustration of plain block (left) and residual block (right).

2) Inception V3 Archeticture

Alongside with ResNet, we utilize GoogleNet Inception V3 CNN architecture [7]. Inception V3 is a well-documented network that is based on inception modules. Inception blocks consist of a series of convolutions in parallel and with different kernel sizes to extract features and finally aggregate the results as shown in Fig 3. The input image is projected to sequence of convolutional and pooling layers, then stack of inception modules for feature extraction. The classifier part in inception V3 is equipped with dropout layer to reduce overfitting [17], and softmax output layers. Again, all convolutions are followed by ReLU as an activation function.

C. Training

Our model was implemented using PyTorch deep learning framework. Adam optimization algorithm was used for neural work optimization. We adapted a dynamic learning rate which was divided by 10 every 30 epochs with an initial value of $5x10^{-5}$ and $1x10^{-4}$ in Inception V3 and ResNet experiments respectively. We used pretrained weights on ImageNet for the convolutional blocks as an initialization for the network parameters, which resulted in a faster convergence (in less than 100 epochs). The training batch size was 16 images for both models.

We used weighted cross entropy which is a modified formula of the well-known loss function - cross entropy, to alleviate the effect of dataset imbalance. Weighted cross entropy takes the following form:

$$H(p, q) = -w_i \sum_x p(x) \log q(x) \quad (1)$$

Where: $p(x)$ is the ground truth label, $q(x)$ is the predicted softmax probability of the neural network, and w_i is the weight for class i.

D. Data Augmentation

Data augmentation module has been added to our workflow to train the network with different variation of the input images by artificially enlarged training set. Data augmentation effect has been practically demonstrated to reduce network overfitting, and consequently help the model to generalize properly.

In our experiments, the selected augmentation methods were horizontal flipping with probability of 0.5, vertical flipping with another probability of 0.5, image rotations with probability of 0.8 using a random angle in the range [-25, 25], and random zooming factor in the range of [0.7, 1.3] with probability of 0.8.

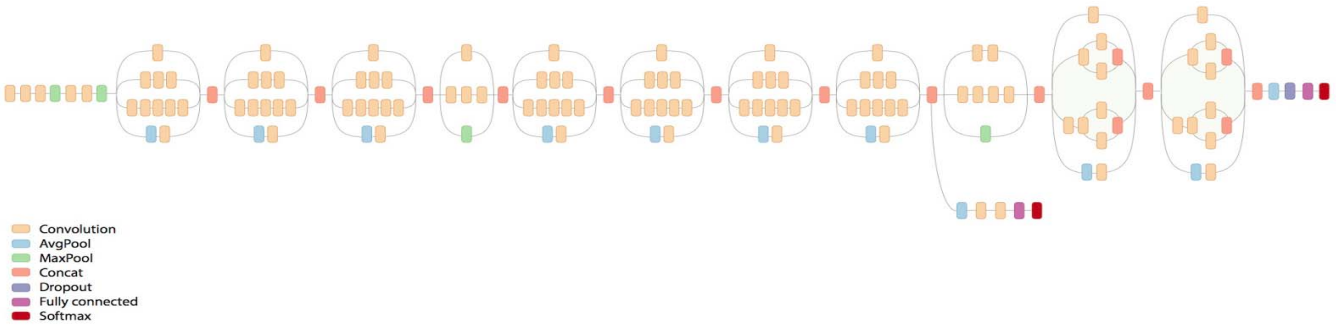

Fig. 3. Inception V3 CNN architecture (reproduced from [18]).

IV. Experimental Results

In our experiments, data was split into 80% as a training set (8012 images) and 20% as a validation set (2003 images). The validation set was used to choose the best epoch for validation score. we report the validation results of standalone ResNet, standalone Inception V3, and ensemble of the two approaches. Eensemble is calculated by averaging the output probabilities of the two networks and selecting the class with the highest probability. We evaluate our model based on accuracy, average precision, and average recall. Our results on the validation set are reported in Table. 1 and Fig. 4.

TABLE I. VALIDATTION RESULTS OF THE THREE APPROACHES

Method	Accuracy	Average Precision	Average Recall
Lopez et al. [13]	80.7%	0.701	0.615
Standalone ResNet-50	87.1%	0.786	0.77
Standalone Inception V3	89.7%	0.849	**0.8**
Esnsemble	**89.9%**	**0.862**	0.796

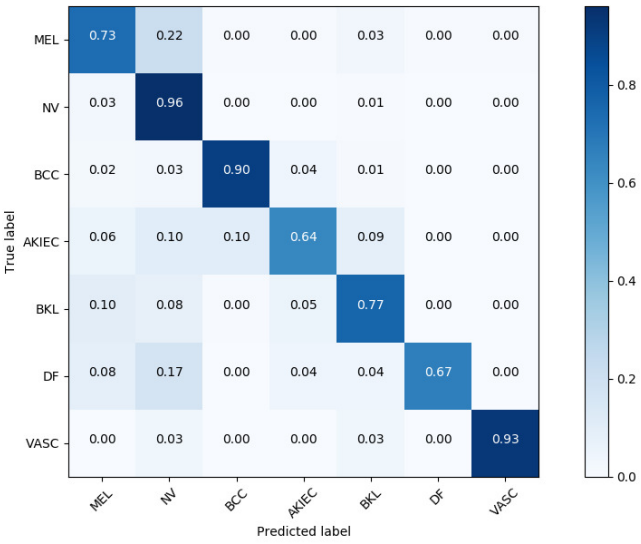

Fig. 4. The confusion matrix of the ensemble classification accuracy (ResNet-50 and Inception V3 combined). MEL: Melanoma, NV: Melanocytic Nevus, BCC: Basal Cell Carcinoma, AKIEC: Actinic Keratosis, BKL: Benign Keratosis, DF: Dermatofibroma, and VASC: Vascular Lesion.

V. Discussion

Automatic diagnosis of melanoma lesions is an important step towards increasing the prevalence of recovery for melanoma patients. In this work, we proposed a convolutional neural network-based framework for lesion classification to seven classes of lesions. We tested two widely-used architectures in visual classification tasks ResNet and Inception V3. To the best of our knowledge, no studies have proposed the usage of ensemble of these architectures for the problem of skin lesions classification.

While inception V3 had a superior performance in most of the images, It was not surprising to notice that taking ensemble of the two architectures by averaging the output probabilities has leveraged the accuracy. Ensembles allowed to get a correct class for some challenging images by promoting the vote of the network that is more confident about its decision (i.e. has a higher probability). Image samples for improvements introduced when used ensemble method for classification are shown in Fig. 5, where the individual classification of each method for the images are presented. We can see that ensemble approach helped identifying the correct class, especially between the two confusing classes (AKIEC and BCC).

We used pretrained weights on ImageNet as an initialization for networks parameters. This speeded up the convergence compared with training from scratch. However, the idea of transfer learning (freezing the convolutional layers with weights trained on ImageNet) failed to get an acceptable performance in our problem, due to the major differences between the dermoscopic images and ImageNet.

Compared to the previous work of Lopez et al. [13], our solution has shown better accuracy. They used VGG network with pretrained weights on ImageNet. They freeze the lower-level layers of the network and train only the top layers. Their classification accuracy on ISIC 2018 dataset was 80.7%. We believe that our solution's superiority is due to: our used deeper architecture, and the ensemble approach.

Our study has suffered from some limitations such as the dataset imbalance, we used weighted cross entropy as a loss function to alleviate its effect. Moreover, due to time and resources constraints we did not manage to run k-fold cross validation experiments, in order to ensure the stability and robustness of our solution.

For the future work, an extensive study is needed to investigate other ensemble techniques, and other CNN classification architectures' (such as: Dense Net, ResNeXt, etc.) performance in skin lesion classification task.

(a) (b) (c)

Fig. 5. Three lesion image examples that have been misclassified by ResNet or Inception deep convolutional neural networks but correctly classified by both networks ensemble. (a) ResNet classification: BCC, Inception classification: AKIEC, ensemble classification: AKIEC, ground truth: AKIEC. (b) ResNet classification: NV, Inception classification: MEL, ensemble classification: MEL, ground truth: MEL. (c) ResNet classification: AKIEC, Inception classification: BCC, ensemble classification: AKIEC, ground truth: AKIEC.

VI. Conclusion

In this work, we introduced our ensemble-based deep learning approach that helps dermatologists robustly differentiate between different seven types of skin lesion categories. We relied on combining two well-known architectures (ResNet-50, and Inception V3) for lesion classification. Our ensemble approach using these networks have shown significantly higher results compared to previous lesion classification approaches. Our proposed approach does not require any prior segmentation, pre-processing or grayscale conversion.

Acknowledgements

In memory of Mohamed Samy Ahmed Kassem (September 1993 – May 2018).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA. Cancer J. Clin.*, vol. 68, no. 1, pp. 7–30, Jan. 2018.

[2] F. Nachbar *et al.*, "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions.," *J. Am. Acad. Dermatol.*, vol. 30, no. 4, pp. 551–9, Apr. 1994.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] E. Nasr-Esfahani *et al.*, "Melanoma detection by analysis of clinical images using convolutional neural network," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 1373–1376.

[5] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[8] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[9] C. Barata, M. Ruela, M. Francisco, T. Mendonca, and J. S. Marques, "Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014.

[10] N. Codella *et al.*, "Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images," vol. 61, no. 4, pp. 1–28, 2016.

[11] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions," *J. Biomed. Inform.*, vol. 34, no. 1, pp. 28–36, Feb. 2001.

[12] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images," 2015, pp. 118–126.

[13] A. Romero-Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, "Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques," in *Biomedical Engineering*, 2017.

[14] D. Gutman *et al.*, "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv Prepr. arXiv 1605.01397*, May 2016.

[15] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172.

[16] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, p. 180161, Aug. 2018.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[18] "Google AI Blog: Train your own image classifier with Inception in TensorFlow." [Online]. Available: https://ai.googleblog.com/2016/03/train-your-own-image-classifier-with.html. [Accessed: 20-Aug-2018].